Causal Inference

SSE 708: Machine Learning in the Era of Big Data

David McCoy PhD, MSc
Division of Biostatistics, UC Berkeley
Edward's Lifesciences

July 15-19, 2024

Table of Contents

What is special about Causal Inference?

- Data + statistical assumptions = statistical inference.
  - → Conclusions about an underlying population.
- Data + statistical assumptions + causal (non-testable) assumptions = causal inference.
  - → Conclusions about how the underlying population would change if conditions changed, e.g, if we changed the way treatment was assigned.
- It also provides scientifically (or policy-wise) interesting quantities in the absence of a parametric model, e.g., not having the convenient coefficients in a model like $E(Y|X) = \frac{1}{1+exp(-(b_0+b_1X_1+\cdots))}$

Roadmap for Estimation

1. Specify a question, causal model, and its link to the observed data: e.g., impact of treatment on disease, $W \rightarrow A \rightarrow Y$.
2. Specify the causal quantity of interest: e.g., $E(Y(tx) - Y(nottx))$.
3. Assess identifiability: e.g., $E(Y(tx)) = E_W[E(Y|A = tx, W)]$
4. Commit to a statistical model and target parameter of the observed data distribution: $O = (W, A, Y) \sim P_0 \in \mathcal{M}$
5. Estimate the chosen parameter of the observed data distribution: $Ave\{\hat{E}(Y|A = tx, W) - \hat{E}(Y(|A = nottx, W)\}$
6. Interpret results

Example: Does Physical Therapy Reduce Hospital Re-admission Rates among COPD patients

- Use the Norwegian health database as an example.
- Non-Causal
  - → How accurate a prediction algorithm for hospital re-admission can we create from the data?
  - → What variables are most important for predicting hospital re-admission for COPH patients?
  - → What are the adjusted odds ratios of yes/no for different utlizations?
- Causal
  - → How would increasing the recent past utilzation of physical therapy impact hospital admission for COPD patients?
  - → What longitudinal pattern of physical therapy appointments minimize the risk of hospital admission?

Example: Transfusion Ratios and Outcomes of Severe Trauma

- In trauma units, patients get reconstituted blood (plasma, RBC, platelets), but what ratios these should be given has been open to debate
  - → Different analyses and trials have suggested different ratios.
- Causal question 1: What ratio of products given to *all* patients (regardless of initial injury, etc.) would result is lowest mortality?
- Causal question 2: What ratio of products given to *specific* patients (based upon of initial injury, etc.) would result is lowest mortality risk for that patient?

Causal Model

- Causal Model is a way to represent background knowledge about the system you want to study Pearl2000
- Example:
  - → What factors affects a patient's attending physical therapy?
  - → What are major determinants of hospital re-admission in the population of interest?
- Structural Causal Models (SCM) unify structural equations, causal graphs, and counterfactual frameworks
- SCM are equations that relate a variable to its causes (parents).

Structural Causal Models: Motivation

- Provide a framework in which we can
  1. express causal assumptions,
     - → these are different from statistical assumptions,
  2. defines a theoretical experiment that defines the causal question,
     - → If everyone with a previous hospitalization from complications of COPD were provided physical therapy, how would the rate of hospital re-admission change?
  3. evaluates whether the data and assumptions are sufficient to answer those questions.
- Once we have succeeded in defining our question as a parameter of the observed data distribution, we are back in the world of standard statistics.

Practice Problem One in Causal Thinking

Use aggregate or the segregated data?

1. There are two treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore, more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data or the stone size-specific data when determining which treatment will be more effective?

2. There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery, and the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately to maximize the chance of a successful surgery?

[1] Pearl, Judea, et al. Causal Inference in Statistics: A Primer, John Wiley & Sons, Incorporated, 2016.

Practice Problem Two in Causal Thinking

A baseball batter Tim has a better batting average than his teammate Frank. However, someone notices that Frank has a better batting average than Tim against both right-handed and left-handed pitchers. How can this happen? (Present your answer in a table.)

Definition of Structural Causal Models (SCM)

1. Endogenous variables.

- Variables that are meaningful for the scientific question, or about which you have some scientific knowledge: $X = \{X_1, X_2, \ldots, X_J\}$
  - → We often (but not always) know the time ordering of these variables
  - → Includes all the variables you measure (or are considering measuring)
  - → Might also include some variables you do not/cannot observe (sometimes called *latent variables*)

SCM's continued

2. Exogenous variables (unmeasured errors):
   $U = \{U_1, U_2, \ldots, U_J\}$.

- Given we never predict things perfectly, need a formal way to incorporate error.
- Not affected by other factors in the model
- All the unmeasured factors not included in X that go into determining the values that the X variables take
  - → U collapses all these unknown factors into one variable
- We denote the distribution of these factors P_U

SCM's continued

.

3. Functions $F = \{f_{X_1}, f_{X_2}, \ldots, f_{X_J}\}$ defines a set of structural equations for each of the endogenous variables.
4. They are mechanism by which each variable is determined given the combination of exogenous and endogenous variables that impact the variable.

- For each endogenous variable in X_j, we specify its parents $P_a(X_j)$. Endogenous variables that may affect the value of X_j:

$$X_j = f_{X_j}(P_a(X_j), U_{X_j}), j = 1, \ldots, J$$

- One common option in the absence of more specific information: include in $P_a(X_j)$ all variables that temporally precede X_j.

SCM's continued

.

- Given an input, the functions f's deterministically assign a value to each of the endogenous variables
- If you know the f's, you would know the data-generating distribution of your variables and all questions could be addressed by these functions
- In truth, we seldom know these so they must be estimated by data
- Our model says that the distribution of (U, X) is generated by
  1. Drawing a multivariate U from a specific probability distribution P_U
  2. Deterministically assigning X by plugging U into the set of functions F
- A given input gives us a specific realization x for each variable in the model.

SCM Encode Causal Assumptions

.

- Assumptions about how the variables X were generated in the system we want to study
- What factors does "nature" (or the "experiment" that generated the data in the system we want to study) consult when assigning a value to these variables?
  - → What do we know about factors determining whether an individual gets a heart attack?
  - → What do we know about factors that affect whether an HIV patient is prescribed abacavir?

Revisit Transfusion Ratios and Outcomes of Severe Trauma

- **Question**: Does the use of platelets in acute trauma decrease mortality due to lack of hemostasis?
- Introducing notation in the single time point treatment context (say any platelets within the first 1/2 hr of admission).
  - → A: indicator of platelets $(0 = no, 1 = yes)$.
  - → W: vector of patient characteristics at baseline before admission to ER.
    - age, type of injury, heart rate, blood pressure, ...
  - Any relevant mediators?
  - → Y is the indicator of death within the next 6 hours $(0 = no, 1 = yes)$.

How a question becomes a parameter of interest (target parameter)

Question - what experiment would one do, if you could do any experiment (regardless of ethics and practicality) to estimate the impact of platelets?

SCM for Point Treatment - the notation and simulation

- $X = \{W, A, Y\}$
- Errors: $U = (U_W, U_A, U_Y)$
- Structural Equations (nature):
  - → $W = f_W(U_W)$
  - → $A = f_A(W, U_A)$
  - → $Y = f_Y(W, A, U_Y)$

Distribution of (U, X) generated by:

1. Draw U from P_U
2. Generate W via f_W with input U_W.
3. Generate A via f_A with inputs (W, U_A)
4. Generate Y via f_Y with inputs (W, A, U_Y)

Nonparametric Structural Equation Models (NPSEM)

- The structural equations do not restrict the functional form of the causal relationships they specify
  - → Eg., specify $Y = f_A(W, A, U_Y)$ rather than $Y = \beta_0 + \beta_1 A + \beta_2 Age + \beta_3 * BP + \ldots + U_Y$
  - → If you have real knowledge about the functional form of a structural equation, you can incorporate it, but generally this is rare in public health.
- Similarly, we do not impose unsupported assumptions on the error distributions (P_U)
- The use of non-parametric structural equation models allows us to respect the limits of our knowledge

Assumptions on the SCM

1. Exclusion Restrictions
   - We make Assumptions by leaving X variables out of a given parent set.
     - → Excluding a variable from $P_a(X_j)$ assumes it does not directly affect what value X_j takes
     - → Leaving a variable in $P_a(X_j)$ just means it might affect what value X_j takes:

$$X_j = f_{X_j}(P_a(X_j), U_{X_j}, j = 1, \ldots, J$$

   - One option: include in $P_a(X_j)$ all variables that temporally precede X_j.

2. Independence Assumptions
   - Independence Assumptions restrict the allowed distributions for P_U
   - E.g., Assume U_A is independent of U_Y:
     - → Corresponds to saying that A and Y share no common causes outside other than those included in X.
     - → Reasonable?

Goals

- Whenever possible, restrict our assumptions to those supported by our knowledge
- When we have to make more questionable assumptions
  - → Make them explicitly so that we can evaluate them better and keep our interpretations honest
  - → If little is known about statistical model, limit them to (causal) assumptions that do not constrain the statistical model

Causal Graphs

- Connect parents to children with an arrow
  - → Makes the asymmetry of the equations explicit
- Each endogenous X variable has an error (U)
- Correlations between errors encoded in dashed lines/double headed errors.

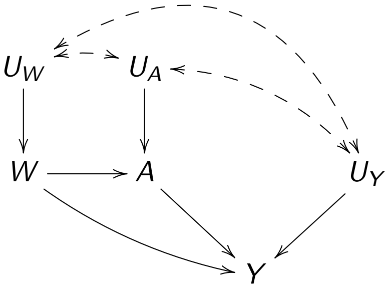


Figure: A causal graph for (1) with no assumptions on the distribution of P_U

Defining a Target Causal Quantity (DaTCQ)

- Recall our motivation: experimental conditions under which we observe a system $\neq$ experimental conditions in which we are most interested
- The process of translating our background knowledge into a SCM required us to be specific about our knowledge of existing experimental conditions.
- The process of translating our scientific question into a target causal parameter requires us to be specific about our ideal experimental conditions.

DaTCQ, cont.

- **Step 1: Decide upon which variable or variables we want to intervene**
  - → "Exposure" or "Treatment"
  - → Interest in system that modifies the way these variables are generated
  - → Focus on one variable at a single time point
    - Lots of times you are interested in intervening on more than one variable/time point
    - We refer to this variable as the intervention variable, and typically use "A" to represent it

DaTCQ, cont.

- **Step 2: Decide in what kind of intervention you are interested**
  - → We focus on "static" interventions
    - Interventions that deterministically set A equal to some fixed value(s) of interest
  - → Other options, e.g., dynamic interventions: Set treatment in response to the values of other variables.
- **Step 3: Specify an outcome**

Example: Transfusion and Death in Trauma Patients

- Question: Does use of platelets reduce the risk of death by coagulopathy?

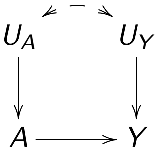
1. What is the intervention variable?
2. What are the interventions of interest?
3. What is the outcome?

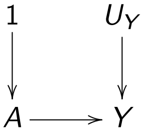
Counterfactuals

- Y_a for an individual is the value that variable Y would have taken for that individual if that individual had received treatment $A = a$
  - → "Counterfactual" because the individual may not have actually received treatment $A = a$
  - → Also referred to as "Potential Outcomes"
- Counterfactuals can be derived from the SCM
  - → If we want to make inferences about data generated by the same system under different conditions, we have to know which parts of the system will change and which parts will stay the same

Interventions on the SCM

- The autonomy of structural equations means that we can make a targeted modification to the set of equations in order to represent our intervention of interest
- Example: intervene on the system to set $A = 1$
  - → Replace f_A with constant function $A = 1$





- $A = f_A(U_A)$
- $Y = f_Y(A, U_Y)$

- $A = 1$
- $Y = f_Y(1, U_Y)$

Target Causal Quantity

1. Decide which variable on which we want to intervene
2. Decide on the intervention(s) of interest

Steps 1 and 2 define our counterfactuals of interest (and our SCM defines a model for the distribution of these counterfactuals)

3. Specify what parameter of the distribution of these counterfactuals of interest (target causal quantity)

Target causal quantity, cont.

- Observed Data (Endogenous Variables)
  - → W = baseline risk factors
  - → A = platelets
  - → Y = hemostasis (or death)
- Structural equations
  - → $W = f_W(U_W)$
  - → $A = f_A(W, U_A)$
  - → $Y = f_Y(W, A, U_Y)$
- Errors: $U = (U_W, U_A, U_Y)\ P_U$
- Full Data: $X^F = (W, (Y_a : a \in \mathcal{A})) \sim \mathcal{F}$ where $\mathcal{A}$ refers to treatment levels of interest, e.g. $\mathcal{A} = \{0, 1\}$.

Example Causal Parameter: Average Tx Effect

- How would expected outcome have differed if everyone in the population had been treated vs. if no one in the population had been treated?
  - → This is a common target of inference.
  - → This is what many RCTs are trying to estimate....

$$ATE = E_{F_X}(Y_1 - Y_0)$$

- Could also have equivalent causal relative risk and causal odds ratio or stratified causal parameter.

Practice Problem 3

Write code in R that can simulate data of the form $O = (W, A, Y)$ from the causal model described above. Make both Y and A binary.

Find the true ATE by simulation.

Marginal Structural Models: When Intervention variable has more than two levels

- Specify a (working) model for $E(Y_a)$ or $E(Y_a|V)$
- Useful when interested in
  - → Dose response curves for multi-level/continuous exposures
  - → Effect modification by multilevel covariates
- Ex: A is number of units of platelets $= (0, 1, 2, \ldots)$

$$
\begin{aligned}
E_{F_X}(Y_a) &= m(a \mid \beta) \\
m(a \mid \beta) &= \beta_0 + \beta_1 a \\
\beta(F_X \mid m) &\equiv \beta E_{F_x}\left[\sum_{a \in \mathcal{A}} (Y_a - m(a \mid \beta))^2\right]
\end{aligned}
$$

The Observed Data from the Actual Experiment

- $O = (W, A, Y)$
  - → Baseline covariates: W = mortality risk factors
  - → Exposure/Treatment: A = platelet use
  - → Outcome: Y = death

- Linking O to SCM
  - → $O = \Phi(X, A)$, where
  - → $X = (Y_1, Y_2, W)$ and
  - → $\Phi(X, A) = (A, W, A * Y_1 + (1 - A) * Y_0)$

The Statistical Model

- The model $\mathcal{M}^{\mathcal{F}}$ (set of possible distributions for (U, X)) implies a model (set of possible distributions) for O
- We refer to this set of possible distributions as the statistical model $\mathcal{M}$
- The true distribution P_0 of O is an element of $\mathcal{M}$
- Often, a model that respects the limits of our knowledge puts no restrictions on the set of allowed distributions for O
- In this case our statistical model is non- parametric and one should respect this fact when framing the statistical estimation problem.
- This generally leads to using data-adaptive, machine learning methods for the necessary regressions.

Positivity in minimal assumption for estimating causal effects

- In plain language: you need enough experimentation of the treatment of interest among people with the same (or very similar) covariates.
- More formally, the relevant assumption is that:

$$\min_{a \in \mathcal{A}} P_0(A = a \mid W) > 0$$

  for all possible W.

- When fitting parametric models, positivity violations are usually invisible. Why? Because such models "extrapolate".

Initial Assumptions Do Not identify Parameter of Interest

- This is more common than not, e.g., in observational studies, always unmeasured confounding.
- Options
  - → Get more background information or data
  - → Give up.
  - → Estimate the association the best way you can with the least residual confounding
- We will be talking about using combination of causal inference and machine learning to do the best job possible.

Practice problem 4

Show on a graph a linear marginal structural model (so EY_a vs. a) which is quadratic in a. Show what would be the estimated difference in counterfactual means if $a = 0$ vs. $a = 3$ with respect to the coefficients of the model.

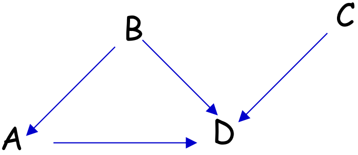
Using counterfactual notation, what would be a causal attributable risk (say $A = 0$ means unexposed)?

Causal Graph Approach to Confounding

- Possible causal effects of childhood vaccination on autism
  - access to general medical care may affect autism incidence and/or diagnosis
  - Access to medical care increases vaccination
  - Family SES influences access to medical care and also ability to pay for vaccination
  - Family medical history may affect risk for autism and may also influence access to medical care
- Which of medical care access, SES, and family history are confounders? Do we need to stratify on all three?

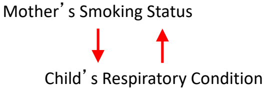
Directed Acyclic Graphs

- Nodes, directed graphs (edges have direction)

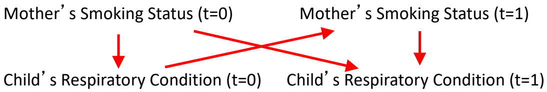


- Directed paths: B-A-D ✓
  B-D-A, C-B-D ☒

Directed Acyclic Graphs

- Acyclic
  - No loops: A cannot cause itself
  - 
    Mother’s Smoking Status

    Child’s Respiratory Condition



Mother’s Smoking Status (t=0)

Mother’s Smoking Status (t=1)

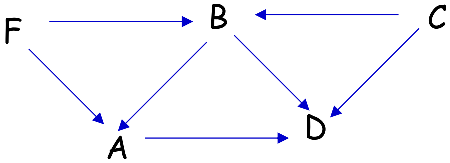
Child’s Respiratory Condition (t=0)

Child’s Respiratory Condition (t=1)

  - node *A* at the end of a directed path starting at *B* is a *descendant* of *B* (*B* is an *ancestor* of *A)*

Directed Acyclic Graphs

- A node A can be a collider on a specific pathway if the path entering and leaving *A* both have arrows pointing into A. A path is blocked if it contains a collider.

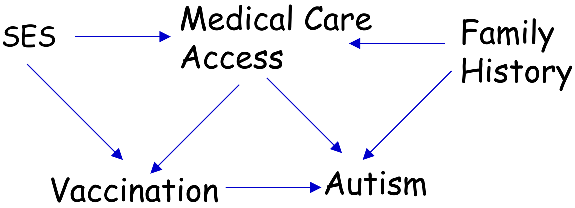


- *D* is a collider on the pathway *C-D-A-F-B;* this path is blocked

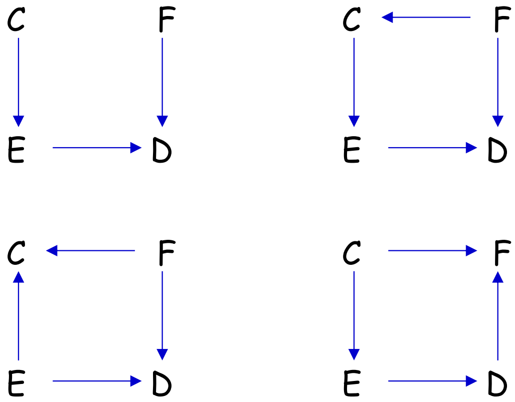
Using Causal Graphs to Detect Confounding

- Delete all arrows from *E* that point to any other node
- Is there now any unblocked backdoor pathway from *E* to *D*?
  - Yes—confounding exists
  - No—no confounding

Vaccination & Autism Example



Using Causal Graphs to Detect Confounding

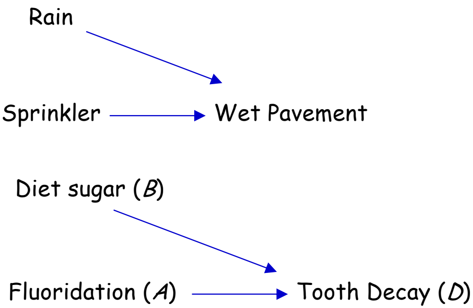


Checking for Residual Confounding

- After stratification on one or more factors, has confounding been removed?
  - Cannot simply remove stratification factors and relevant arrows and check residual DAG
  - Have to worry about colliders

Controlling for Colliders

- Stratification on a collider can induce an association that did not exist previously



Hypothetical Data on Water Flouridation, High-Sugar Diet, and Tooth Decay

			Tooth Decay (D)			
Flouridation			D	$\overline{D}$	OR	ER
A (no fluoridation)	High-Sugar Diet	B	160	40	2.67	0.2
		$\overline{B}$	120	80		
$\overline{A}$ (fluoridation)	High-Sugar Diet	B	80	120	2.67	0.2
		$\overline{B}$	40	160		
			Tooth Decay (D)			
High-Sugar Diet			D	$\overline{D}$		
B	Fluoridation	A	160	40	6.00	0.4
		$\overline{A}$	80	120		
$\overline{B}$	Fluoridation	A	120	80	6.00	0.4
		$\overline{A}$	40	160		
Pooled table			Fluoridation			
			A	$\overline{A}$		
	High-Sugar Diet	B	200	200	1.00	0.00
		$\overline{B}$	200	200		

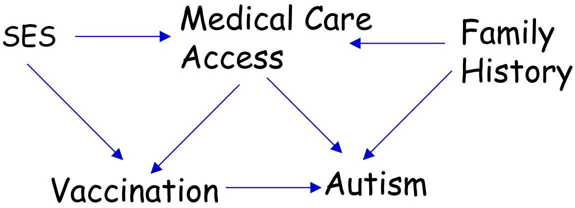
Hypothetical Data on Water Flouridation, High-Sugar Diet, and Tooth Decay

Tooth Decay (D)					
		Fluoridation			
		A (no flouridation)	$\overline{A}$ (no flouridation)	OR	ER
	B	160	80	0.67	-0.083
	$\overline{B}$	120	40		
No Tooth Decay ($\overline{D}$)		Fluoridation			
		A (no flouridation)	$\overline{A}$ (no flouridation)	OR	ER
	B	40	120	0.67	-0.083
	$\overline{B}$	800	160		

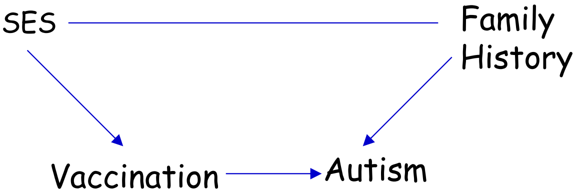
Checking for Residual Confounding

- Delete all arrows from *E* that point to any other node
- Add in new undirected edges for any pair of nodes that have a common descendant in the set of stratification factors *S*
- Is there still any unblocked backdoor path from *E* to *D* that doesn’t pass through *S* ? If so there is still residual confounding, not accounted for by *S* .

Vaccination & Autism Example

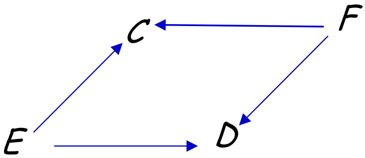


Vaccination & Autism Example: Stratification on Medical Care Access



Still confounding: need to stratify additionally on SES or Family History, or both

Caution: Stratification Can Introduce Confounding!



No Confounding

Stratification on *C* introduces confounding!

References

J. Neyman.
On the application of probability theory to agricultural experiments (1923).
Statistical Science, 5:465–480, 1990.

J.M. Robins.
The analysis of randomized and non-randomized aids treatment trials using a new approach in causal inference in longitudinal studies.
In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Methodology: A Focus on AIDS*, pages 113–59. U.S. Public Health Service, National Center for Health SErvices Research, Washington D.C., 1989.

J.M. Robins.
Causal inference from complex longitudinal data.
In M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer Verlag, New York, 1997.

D.B. Rubin.
Statistics and causal inference: Comment: Which ifs have causal answers.
J. Am. Statist. Assoc., 81:961–2, 1986.